

Leveraging LLMs for Fair Data Labeling and Validation in Crowdsourcing Environments [Vision Paper]

1st Ibrahim MOHAMED SEROUIS
IRIT
Université Toulouse III
Toulouse, France
ibrahim.mohamed-serouis@irit.fr

2nd Florence SÈDES
IRIT
Université Toulouse III
Toulouse, France
florence.sedes@irit.fr

Abstract—The rapid expansion of large datasets, encompassing text, images, audio, and video, presents substantial challenges for data labeling, a crucial step in machine learning and data science workflows. Large Language Models (LLMs) provide a promising approach for automating and improving the accuracy of data labeling across various modalities. However, their application in this area introduces specific challenges, such as managing diverse data types, maintaining high-quality annotations, handling computational complexity, and, critically, addressing and mitigating biases associated with automated methods. This vision paper examines the potential of LLMs to enable fair data labeling and validation in crowdsourcing environments, discussing the current landscape, and existing challenges, providing potential future research directions that can help ensure their successful integration.

Index Terms—Large-scale data labeling, Automated annotation, Large Language Models, Fair data labeling

I. INTRODUCTION

The development of machine learning heavily depends on the advancements in computing power and the widespread availability of annotated data sources, which are crucial for creating datasets that machines can learn from effectively. After the data collection phase, data labeling becomes a critical step, involving the assignment of appropriate tags or labels to datasets. This process is essential for the effective training of artificial intelligence (AI) models, allowing them to recognize and categorize various data types, such as images, audio files, video footage, and text content.

Initially, data annotation was primarily carried out by human annotators. However, this process has significantly evolved with the increased use of crowdsourcing, which distributes tasks across larger groups, and partial automation, driven by recent advancements in AI research. Despite these improvements, the labeling processes are still prone to errors and are likely to perpetuate human biases within the datasets.

Recently, Large Language Models (LLMs) have exhibited impressive capabilities across a wide range of tasks, particularly in understanding and generating human-like text, images, and speech. These capabilities could be leveraged to assist in the labeling of large datasets. However, utilizing

LLMs for large-scale data labeling presents several challenges. These models need to handle the complexities of various data types, ensure accuracy and consistency, manage large-scale data processing, and, importantly, address the inherent biases in the LLMs related to the specific task or problem.

Contributions. This vision paper offers a comprehensive overview of these challenges and proposes a framework for employing LLMs to develop more efficient, fair, and accurate data labeling systems, particularly for large-scale datasets.

Paper structure. In Section II, we provide an overview of related research. In Section III, we detail the challenges associated with integrating LLMs into data labeling workflows, laying the foundation for our proposed vision in Section IV, which we articulate through four key recommendations. In Section V, we examine the challenges related to the practical implementation of our recommendations.

II. BACKGROUND

A. Large Language Models

Currently, Large Language Models (LLMs) represent some of the most significant advancements in artificial intelligence. Numerous variants of LLMs are available, including various versions of GPT ([1]–[4]), XLNet [5], Meta’s Llama [6], and Google’s Gemma [7], with foundational models like Transformers [8] and BERT [9] leading the way in these developments. Beyond excelling in natural language processing tasks, LLMs have demonstrated remarkable adaptability across various fields, including education, dataset generation, healthcare [10], and scientific research [11]. Trained on extensive text corpora, these models are capable of performing a wide range of language and vision tasks, such as translation, summarization, and question answering. Their broad success indicates a promising potential for application in the generation of data labels.

B. Traditional data labeling techniques and crowdsourcing platforms

Traditional data labeling techniques primarily depend on human annotators, making the process time-intensive, costly,

and susceptible to errors. Crowdsourcing platforms such as Amazon Mechanical Turk, CrowdFlower, and Prolific have partly mitigated these challenges by distributing labeling tasks among numerous contributors, thus becoming essential components in some data labeling workflows. While these platforms facilitate the creation of large-scale annotated datasets, they also come with limitations, including inconsistencies in annotator expertise and the introduction of potential biases. Annotators are frequently not trained in the specific domain relevant to the data they are labeling [12]–[14], or may belong to similar demographic groups [15]. Furthermore, biases within crowdsourcing environments are often overlooked as a significant source of low-quality or biased data [16].

C. Efforts to integrate LLMs in data labeling processes

Recent advancements in the literature have begun to explore the integration of Large Language Models (LLMs) into data mining and labeling processes [17], [18], as well as into the generation and labeling of data [19], potentially setting the stage for future research in this area. However, the frameworks proposed in these studies have largely overlooked the ethical and fairness concerns associated with using LLMs for these purposes. While some research, such as [20], has suggested hybrid approaches that combine LLMs with human experts to promote fairer data generation, they fail to adequately address the biases in the labeling process before re-processing and the inherent biases within the LLMs themselves.

III. CHALLENGES IN LLM-DRIVEN DATA LABELING

A. Ensuring Label Accuracy and Consistency

Although Large Language Models (LLMs) offer significant advantages in automating the labeling process, maintaining the accuracy and consistency of the labels they generate is essential. Challenges such as variations in context, data ambiguity, and potential misinterpretations by the model can result in inconsistent or erroneous labels.

B. Managing computational costs

LLMs require substantial computational resources, particularly when applied to large datasets. The cost of running these models can be prohibitive, especially when processing complex, multimodal data at scale.

C. Mitigating bias and ensuring trust

LLMs are known to reflect and amplify biases present in their training data. When applied to data labeling, these biases can lead to skewed datasets, affecting model fairness and performance. Addressing and mitigating these biases is essential to ensure equitable outcomes. Also, building trust among users in the accuracy and reliability of LLM-generated labels is crucial. Users must be confident that the automated labels are as reliable as human-generated labels.

IV. VISION: FUTURE DIRECTIONS

As data collection, validation, and labeling processes are mostly disclosed in most studies, our future visions are mostly based on the awareness of the inherent biases in LLM deployment, developed in Section IV-A, the potential of hybrid methodologies combining LLMs and humans for validation (Section IV-B), and a real-time analysis of the biases in annotated data combined to active learning (Section IV-C).

These recommendations are aimed to be applied altogether to fulfill their goal. Figure 1 can help understanding the overall structure of our concept.

A. Bias awareness in LLM deployment

The growing emphasis on equity and fairness in artificial intelligence (AI) has spurred the development of various applications and methodologies to address bias in machine learning (ML) models. Foundational work by Dwork [21] and Feldman [22] introduced mathematical frameworks and metrics for defining and quantifying bias in ML models. While many of these bias metrics—such as Demographic Parity Difference, Equalized Odds, and Equal Opportunity Difference—are applied to the outcomes of AI models or the datasets they are trained on, these metrics are often not reported when deploying or publishing technical reports on Large Language Models (LLMs).

For LLMs trained on open-source data, we propose that biases, particularly those concerning protected groups relevant to the specific tasks for which the models are trained, should be thoroughly investigated and disclosed to the public, either before deployment or retrospectively. This would enable the selection of models not only based on raw performance metrics (such as accuracy, and token generation capabilities in terms of time and memory) but also in terms of bias, ensuring fairness tailored to the specific application. If such data are not disclosed, there can be a systematic audit of the initial selections of LLMs for labeling, beforehand; however, this audit can reveal itself costly in terms of expertise, time, and resources.

B. Hybrid data labeling and validation pipelines

Hybrid approaches that combine the multi-domain capabilities of Large Language Models (LLMs), particularly the larger ones, with human cognitive insights and potential expertise present a promising strategy for both labeling and validating data.

We propose that, as detailed in Section IV-A, employing the fairest (as in the least biased) LLMs as validators within a framework similar to that in [18] can yield both high-quality and fair outcomes. In this approach, both human annotators and LLMs assess the same data points, providing independent ratings of the labels. Discrepancies between these ratings can be analyzed by an automated system using relevant metrics, such as Kappa statistics [23], or reviewed by a domain expert to evaluate the (dis)agreement and resolve any conflicts in labeling. These resulting labels can then be fed to the active learning pipeline mentioned in Section IV-C if it appears to

be a recurrent mistake made by the model, or simply corrected by the expert to generate the correct label.

C. Metrics-aware data labeling processes and active learning

In most studies, bias metrics in labeled data are evaluated post hoc, although some quality metrics, such as completeness (i.e., whether all relevant data points are annotated), are sometimes monitored in real time. However, incorporating bias metrics related to the annotated data in real-time, and pausing the labeling process after a certain number of epochs, can provide insights into both the suitability of the Large Language Model (LLM) for the task and the biases or misinterpretations that may arise from the model's guidelines.

Additionally, establishing and running an active learning pipeline, as defined by [24], can enhance the LLM's annotation capabilities in the specific context of the data. Drawing from the works of [25]–[27], we propose that, by validating data at regular intervals (e.g., every k steps), we can fine-tune the LLM on examples flagged by validators (as discussed in Section IV-B) that initially exhibited biased representations. This iterative process can help ensure greater consistency and fairness in the annotations. This approach is particularly valuable given that, as some studies have shown, even LLMs with extensive multi-domain capabilities may struggle to effectively capture context, especially in nuanced instances [28], [29].

D. Disclosing the overall process

We also contend that transparently disclosing the methodologies outlined in Sections IV-A, IV-B and IV-C can aid researchers and practitioners in better identifying potential sources of bias within these processes. This transparency allows for the development of improved or more appropriate alternatives. Additionally, such openness can theoretically enhance user trust in datasets that adhere to these recommended practices or equivalent implementations.

Users would then be informed of how their data were annotated, what were the initial biases of the LLMs used for the specific application, how the biases were addressed when encountered during the data labeling process, and how a label was deemed correct. Trust would theoretically emerge from transparency in the whole process, which is not always present in the literature.

V. DISCUSSION: PRACTICAL IMPLEMENTATION

In this section, we discuss the key challenges related to the practical implementation of our recommendations.

A. Best application framework

While our recommendations provide valuable insights into potential approaches for integrating Large Language Models (LLMs) into data labeling pipelines, it is important to recognize that their applications are most effective with large datasets, particularly during the data validation phase at each k step. Small volumes of data would not necessarily benefit from halting the labeling process repetitively or fine-tuning on a few instances.

B. Disclosing the overall process: realities

Disclosing the overall process is beneficial in open research, especially within academic settings. However, the *publish-or-perish* culture [30] and the pressure to produce state-of-the-art results can create a competitive environment. This competitiveness may lead some research groups to withhold certain methodological details to avoid being outperformed, although we may argue that it may already be the case in the current research landscape.

C. Computing resources

The computational resources required for training and running inferences with LLMs are substantial. Small organizations and researchers with limited access to computational resources may encounter difficulties in effectively utilizing LLMs for any application. However, advancements in LLM quantization and reduction techniques, as explored by Yao et al. [31], offer promising prospects for making LLM inference and training more accessible to a broader range of researchers.

As things evolve rapidly, especially in this domain, the supercomputers of today that can generate hundreds of tokens in a matter of seconds can be revealed to be the average computers in five years. Inferences that are hard to complete by the LLMs as of now due to context windows and token generation limits can become trivial in just a matter of years. Computing resources should not refrain from the overall interest -for research groups that are able to- in the implementation of those recommendations, as a fairer labeling process can lead to fairer algorithms and may benefit several marginalized groups, especially in sensitive applications that may lead to important decisions made on some users such as crime detection or fraud evaluation, video-surveillance systems, healthcare, credit score attribution systems, or education purposes.

VI. CONCLUSION

The application of Large Language Models to data labeling and validation for large datasets presents both significant opportunities and challenges. By addressing the complexities and challenges related to the inherent bias of the LLMs themselves in their integration to large-scale annotation environments, we can unlock the full potential of LLMs in this domain while ensuring trust by the end-user, and a fairer labeling process. We believe our vision paper outlined a path forward for future research endeavors.

However, for our recommendations to be applicable, the efforts at each level or recommendation should come from the whole machine learning and big data community altogether.

ACKNOWLEDGMENTS

This work has been supported by the French National Research Agency through the ANR TRACTIVE project ANR-21-CE38- 00012-01.

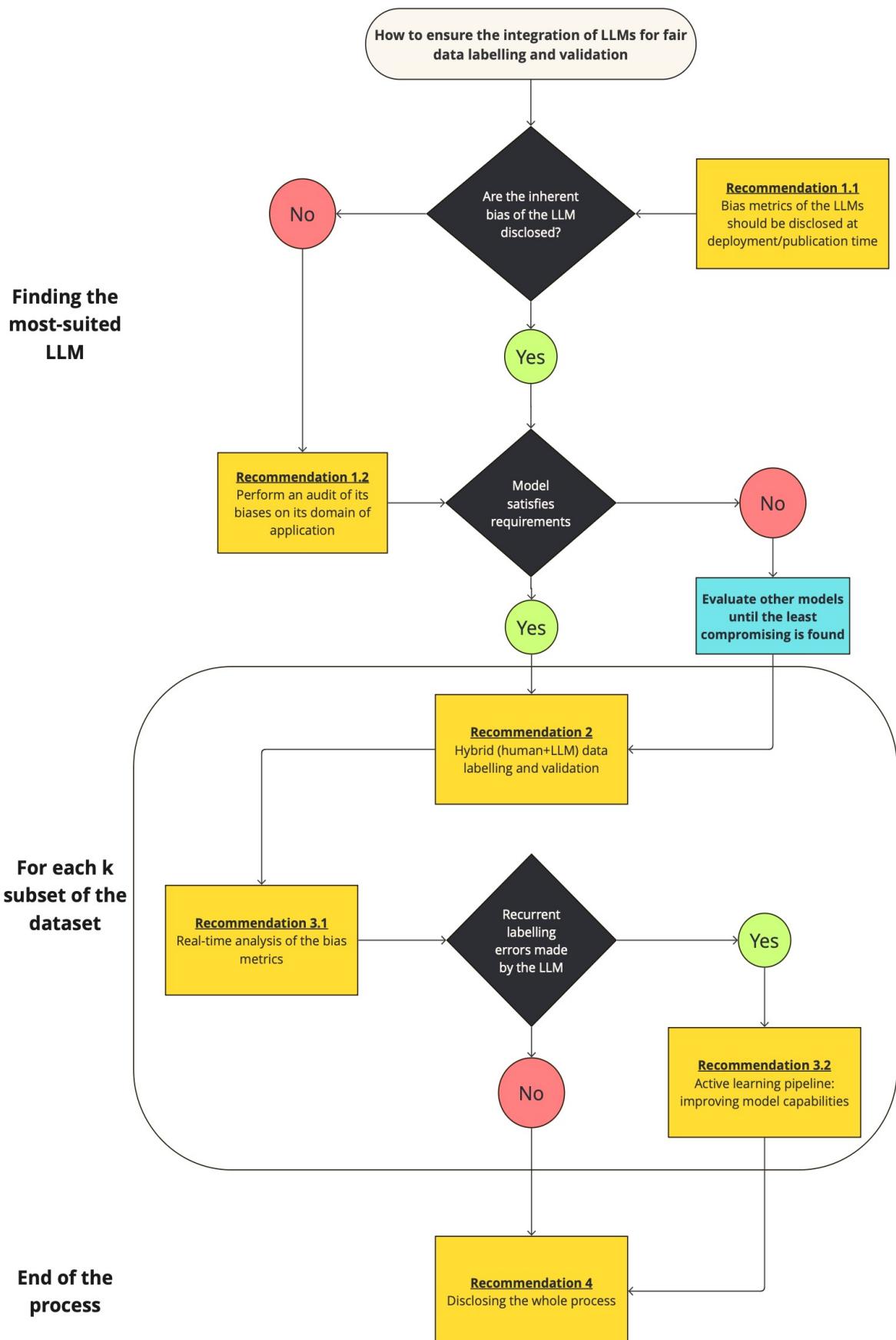


Fig. 1. Diagram of the recommendations. Each recommendation is described in a yellow box.

REFERENCES

- [1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, and others, “Improving language understanding by generative pre-training,” 2018. Publisher: OpenAI.
- [2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, and others, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [3] I. D. Raji, “Handle with Care: Lessons for Data Science from Black Female Scholars,” *Patterns*, vol. 1, no. 8, p. 100150, 2020.
- [4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, and others, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [5] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Advances in neural information processing systems*, vol. 32, 2019.
- [6] H. Touvron, T. Lavigil, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, and others, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [7] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Riviére, M. S. Kale, J. Love, and others, “Gemma: Open models based on gemini research and technology,” *arXiv preprint arXiv:2403.08295*, 2024.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Å. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *North American Chapter of the Association for Computational Linguistics*, 2019.
- [10] M. Sallam, “ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns,” in *Healthcare*, vol. 11, p. 887, MDPI, 2023. Issue: 6.
- [11] A. Jungherr, “Using ChatGPT and other large language model (LLM) applications for academic paper assignments,” 2023. Publisher: Otto-Friedrich-Universität.
- [12] S. V. Rouse, “A reliability analysis of Mechanical Turk data,” *Computers in Human Behavior*, vol. 43, pp. 304–307, 2015.
- [13] A. M. Mellis and W. K. Bickel, “Mechanical Turk data collection in addiction research: Utility, concerns and best practices,” *Addiction*, vol. 115, no. 10, pp. 1960–1968, 2020. Publisher: Wiley Online Library.
- [14] M. A. Webb and J. P. Tangney, “Too good to be true: Bots and bad data from Mechanical Turk,” *Perspectives on Psychological Science*, p. 17456916221120027, 2022. Publisher: SAGE Publications Sage CA: Los Angeles, CA.
- [15] P. G. Ipeirotis, “Demographics of mechanical turk,” 2010. Publisher: NYU working paper no. CeDER-10-01.
- [16] T. Draws, A. Rieger, O. Inel, U. Gadidaju, and N. Tintarev, “A checklist to combat cognitive biases in crowdsourcing,” in *Proceedings of the AAAI conference on human computation and crowdsourcing*, vol. 9, pp. 48–59, 2021.
- [17] M. Wan, T. Safavi, S. K. Jauhar, Y. Kim, S. Counts, J. Neville, S. Suri, C. Shah, R. W. White, L. Yang, and others, “Tnt-llm: Text mining at scale with large language models,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5836–5847, 2024.
- [18] X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao, “Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI ’24, (New York, NY, USA), Association for Computing Machinery, 2024. event-place: Honolulu, HI, USA.
- [19] N. Pangakis and S. Wolken, “Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels,” *arXiv preprint arXiv:2406.17633*, 2024.
- [20] I. M. Serouis and F. Sánchez, “Exploring Large Language Models for Bias Mitigation and Fairness,” in *International Joint Conference on Artificial Intelligence 2024 Workshop on AI Governance: Alignment, Morality, and Law*, IJCAI 2024 Workshop AIGOV, (Jeju Island, South Korea), Aug. 2024.
- [21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.
- [22] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and Removing Disparate Impact,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, (New York, NY, USA), pp. 259–268, Association for Computing Machinery, 2015. event-place: Sydney, NSW, Australia.
- [23] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012. Place: Croatia.
- [24] B. Settles, “Active learning literature survey,” 2009. Publisher: University of Wisconsin-Madison Department of Computer Sciences.
- [25] V. Kulikov, R. Neychev, and I. Makarov, “Whether Large Language Models Learn at the Inference Stage? Effects of Active Learning and Labelling with LLMs on their Reasoning,” in *International Conference on Analysis of Images, Social Networks and Texts*, pp. 42–53, Springer, 2023.
- [26] B. Yuan, Y. Chen, Y. Zhang, and W. Jiang, “Hide and Seek in Noise Labels: Noise-Robust Collaborative Active Learning with LLMs-Powered Assistance,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10977–11011, 2024.
- [27] J. Liang, L. Liao, H. Fei, B. Li, and J. Jiang, “Actively Learn from LLMs with Uncertainty Propagation for Generalized Category Discovery,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7838–7851, 2024.
- [28] A. Sobieszek and T. Price, “Playing games with AIs: the limits of GPT-3 and similar large language models,” *Minds and Machines*, vol. 32, no. 2, pp. 341–364, 2022. Publisher: Springer.
- [29] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, “Large language models struggle to learn long-tail knowledge,” in *International Conference on Machine Learning*, pp. 15696–15707, PMLR, 2023.
- [30] H. P. Van Dale and K. Henkens, “Intended and unintended consequences of a publish-or-perish culture: A worldwide survey,” *Journal of the American Society for Information Science and Technology*, vol. 63, no. 7, pp. 1282–1293, 2012. Publisher: Wiley Online Library.
- [31] Z. Yao, X. Wu, C. Li, S. Youn, and Y. He, “Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 19377–19385, 2024. Issue: 17.